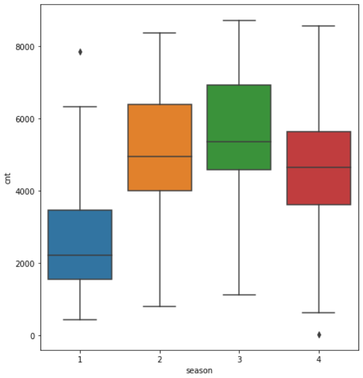
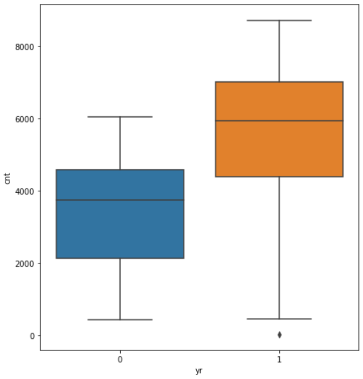
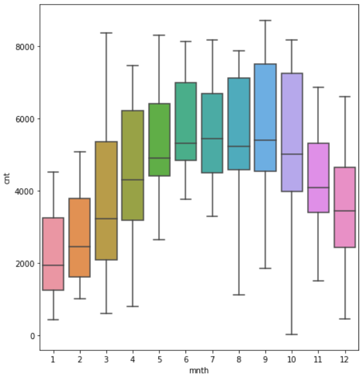
Assignment-based Subjective Questions

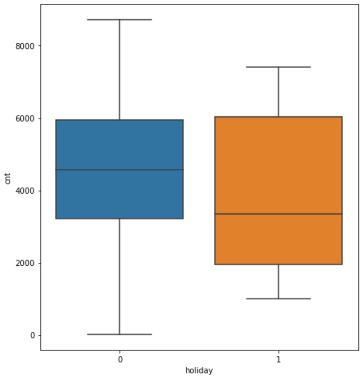
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

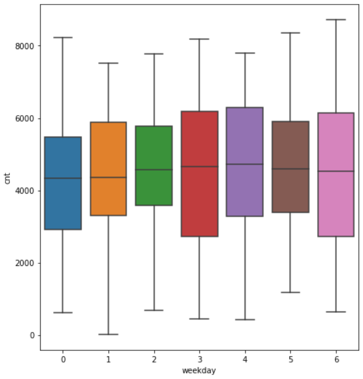
Answer :- The categorical variables of the dataset were analyzed by plotting the boxplots as shown below :-

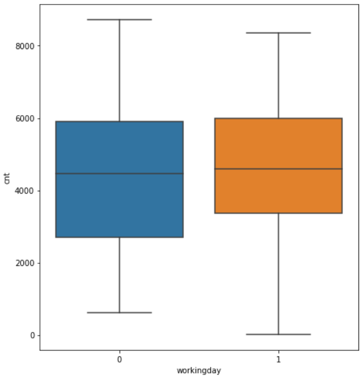


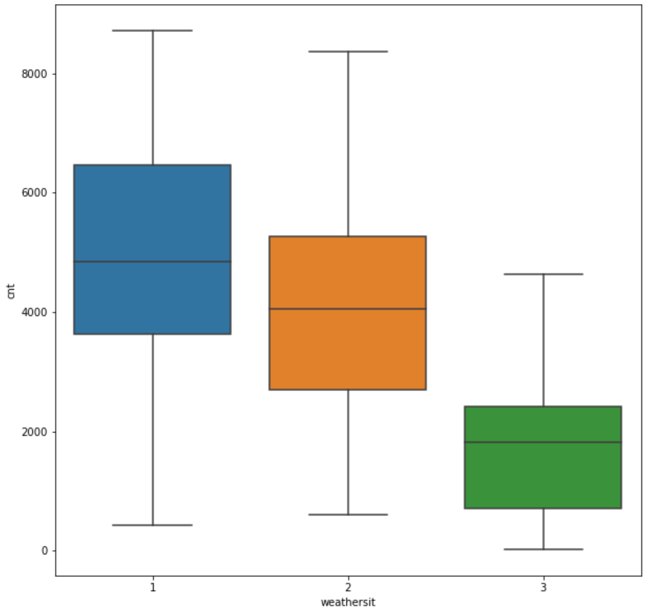












Important observations from above boxplots of categorical variables are as below :-

- season 2(summer) & season 3 (fall) have higher bike bookings. season 1(spring) has the least
- There is a considerable increase in bike bookings from year 2018 to year 2019. Hence, the demand is expected to increase on year by year
- Months 6 to 9 are having most consistent bike bookings, with a median of above 5000 bookings
- Less bookings are observed on holidays, as was expected
- weekdays seem not to have any strong relationship with bookings
- working days are having more consistent and slightly higher bookings than non-working days

2. Why is it important to use drop_first=True during dummy variable creation?

Answer :- drop_first=True, is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

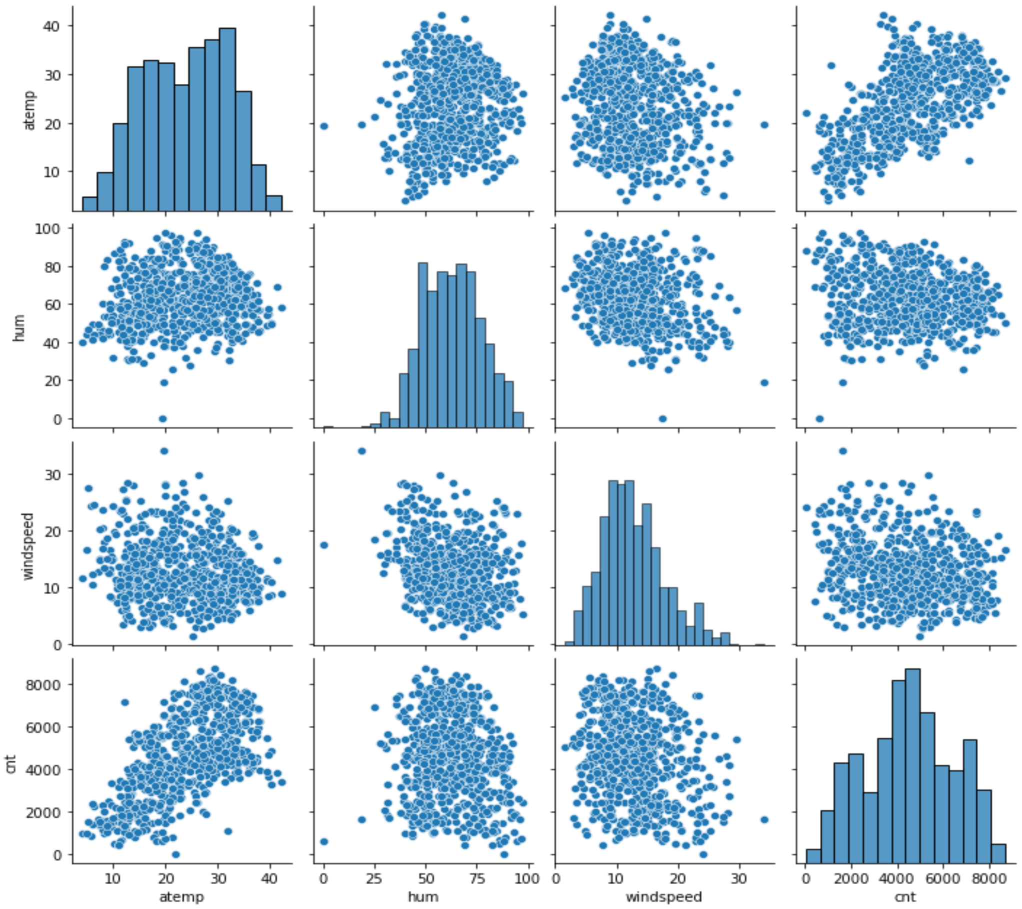
For example, we have created dummy variables for 'season', which had 4 different entries against it. However, only three are required in dummy variables, in order to explain all the 4. As, the dummy variables have representation by binary numbers (0,1). Hence, a '0' in all the 3 selected dummy variables mean that it refers to the 4th condition. The table after dropping 1st column (for season_1) will be as below :-

Season	Season_2	Season_3	Season_4
Season_1	0	0	0
Season_2	1	0	0
Season_3	0	1	0
Season_4	0	0	1

Hence, it can be seen above that season_1 is also explained by the other 3 seasons.

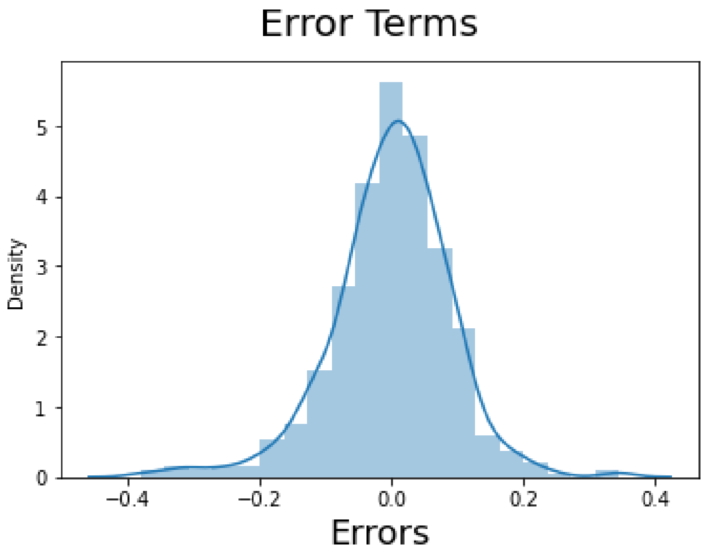
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer :- Looking at the pair-plot (shown below), it looks that the numerical variable **'atemp'** has the highest correlation with the target variable 'cnt'.

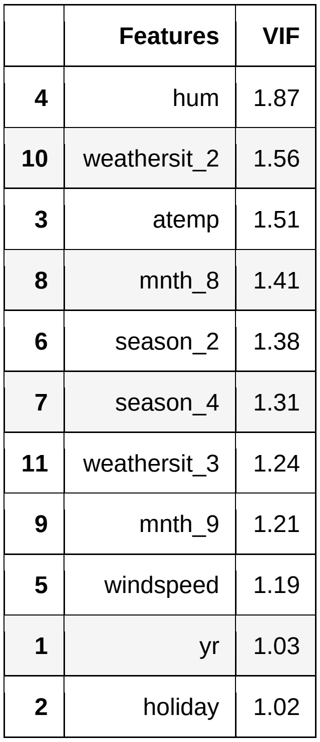


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer :- After building the model on the training set, the assumptions of linear regression was checked through residual analysis of the train data. A histogram of error terms (y_train-y_train_pred) was plotted and it was observed that it follows normal distribution and hence the assumption is correct. The plot is as below :-



Multicollinearity of the independent variables was checked by the VIF values and all values were below “2”, showing that there is no multicollinearity. The results of VIF analysis is as below :-



	Features	VIF
4	hum	1.87
10	weathersit_2	1.56
3	atemp	1.51
8	mnth_8	1.41
6	season_2	1.38
7	season_4	1.31
11	weathersit_3	1.24
9	mnth_9	1.21
5	windspeed	1.19
1	yr	1.03
2	holiday	1.02

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer :- The top 3 predictor variables that influences the bike booking are:

i. ***apparent temperature (atemp)*** - A coefficient value of ‘0.5616’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5616 units.

ii. ***Year (yr)*** - A coefficient value of ‘0.2294’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2294 units.

iii. ***Weather Situation 3 (weathersit_3)*** - A coefficient value of ‘-0.2279’ indicated that a unit increase in Weathersit_3 variable decreases the bike hire numbers by 0.2279 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer :- Linear Regression Algorithm is a machine learning algorithm based on supervised learning. It is one of the very basic forms of machine learning, where we train a model to predict the behaviour of the data, based on some variables. In the case of linear regression as we can see that the name suggests linear which means the two variables which are on the x-axis and y-axis should be linearly correlated.

Linear regression can be done between 2 variables only (1 target and 1 independent variable). This is called simple linear regression. Although in real scenarios, there are multiple variables affecting the target variable and that too at a different significance level. The analysis for this case is called Multiple regression.

An example is the current assignment of bike sharing. Here our target variable is the number of bike bookings and we have multiple other variables available from the dataset. We have to find the linear relationship of the independent variables to the target variable. If we arrive at a significant correlation and are able to depict it in a mathematical equation, then it will be really helpful for the company to project there future plans to enhance their growth and profit. Here the idea is to estimate the future daily bookings of bike sharing, based on the historical data, by learning the behaviour or patterns from the historical data.

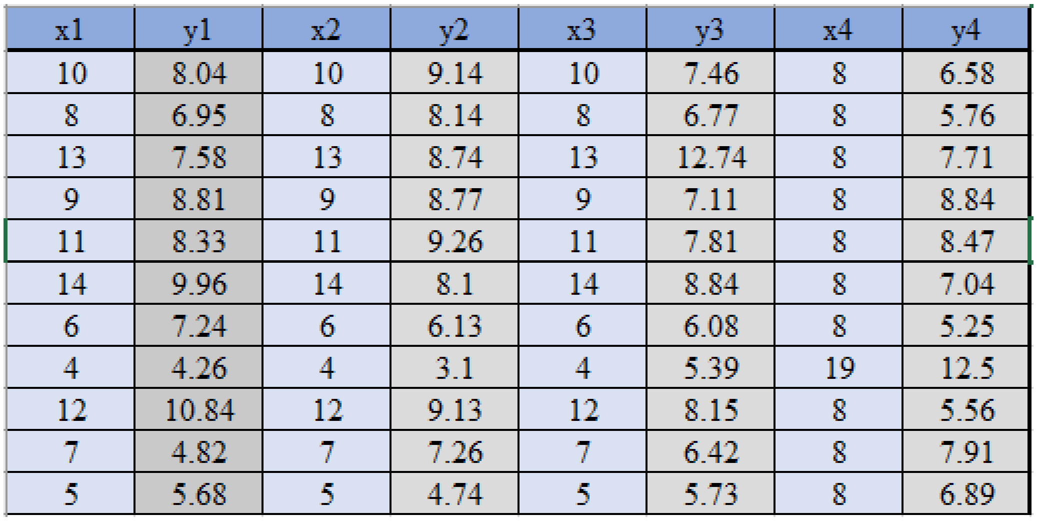
In some cases, the value will be linearly upward that means whenever X is increasing as Y is also increasing or vice versa which means they have a correlation that will be a linear downward relationship.

Linear regression is used to predict the value of target variable Y from the predictor variable X.

2. Explain the Anscombe's quartet in detail.

Answer :- Anscombe's quartet is a perfect example to understand the importance of Data visualization. This was developed 1st by a statistician named Francis Anscombe in 1973. He had taken 4 data sets of X & Y with 11 datapoints of X,Y in each set. All the datasets had almost same basic statistical parameters (Mean, Median, Standard deviation etc.), although when we plot the graphs of X &Y, for the 4 data sets, they look completely different from each other.

The data set considered by Francis Anscombe is as below :-

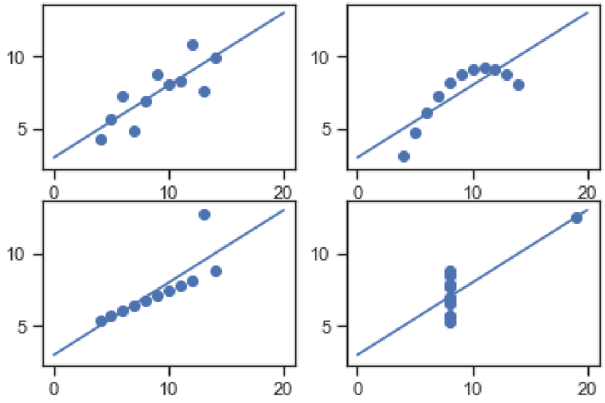


x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

For all the 4 data sets mentioned above, we will get the below mentioned statistics :-

- Xavg = 9
- Yavg = 7.5
- Variance of X = 11
- Variance of Y = 4.12
- Correlation coefficient = 0.816
- Linear regression equation is Y=0.5*X+3

However, when we plot the graphs, it looks like as below :-



Hence, the graphs tell completely a different story about the relationship of X with Y.

Dataset 1 is somewhat linear, Dataset 2 is curved, Dataset 3 is highly linear except a few outliers.

The huge effect of outliers is also very much evident in 4th Dataset, where one value of x4 is 19, whereas all other values are same.

3. What is Pearson's R?

Answer :- Pearson's R is basically the Pearson correlation coefficient, which is a measure of correlation between 2 sets of data. It is very much used in linear regression. Mathematically speaking, it the normalized measurement of covariance. Or in other words, it is the covariance of 2 variables divided by the product of their standard deviations.

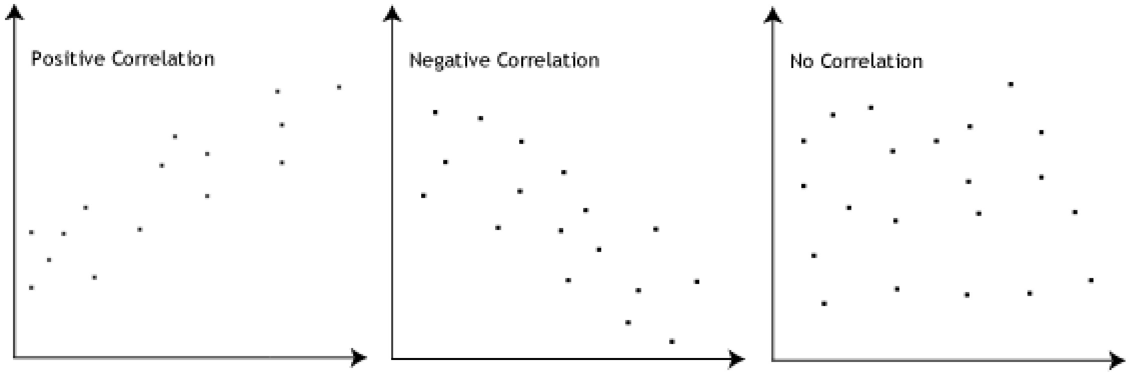
The value always lie between -1 & 1

A value of -1 means that variables are strongly negatively correlated. i.e. with increase in one variable, the other variable decreases

A value of 1 means, that there is very strong +ve correlation. i.e. with increase in one variable, the other also increases.

A 0 value means that there is no correlation between variables.

It can also be seen in below graphs :-



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer :- Scaling is a step followed for independent variables/Features, before analyzing them statistically. This method basically normalizes these variables in a range. Also, it helps in increasing the speed of calculations for the algorithm.

Many a times, the variabes/features provided in the dataset have varying magnitudes, units and range. If scaling on these features is not done, then the prepared model gives importance to the magnitude of variable only, leading to misleading model and interpretation. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. However, it is important to note that scaling affects coefficients of variables only and none of the statistical parameters, such as F-statistics, P-value, R-squared etc. change.

There are two common ways of rescaling:

1. Normalization or Min-Max scaling

2. Standardization (mean-0, sigma-1)

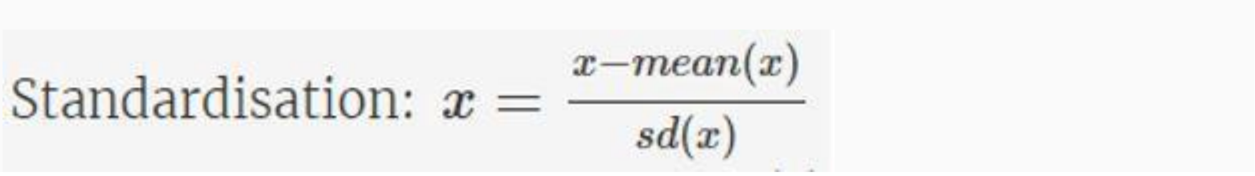
The basic difference between the above scaling methods is as explained below :-

Normalization/MinMax scaling brings all of the data in the range of 0 and 1.

Each value of x is replaced by below formula

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization scaling replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).



Standardisation: $x = \frac{x - mean(x)}{sd(x)}$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer :- we know the formula for VIF is as below :-

VIF = 1/ (1-R^2)

Hence, VIF will be infinite, when denominator is zero. i.e. R^2 = 1.

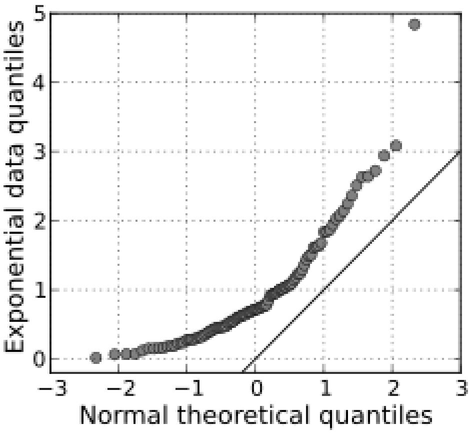
R^2=1 Means that there is a perfect correlation between 2 independent variables. The infinite VIF value of a variable means that all the effect of this variable can be expressed by other variable (as they are having perfect correlation), which also would be having infinite VIF value. To rectify this problem, we should remove one of the variable having perfect correlation with other.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer :- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.